Introduction

Dans le cadre de l'UV IC05, nous avons réalisé un projet de récupération et analyse de données. Notre étude s'est articulée autour de Netflix, célèbre plateforme de streaming. Bien que le site propose au visionnage des films et séries avec au préalable un achat des droits pour les diffuser librement, une certaine quantité des œuvres proposées sont dites « Netflix originals », diffusées exclusivement sur la plateforme. Ainsi, il nous a semblé juste de se questionner quant à la proportion de ces films et séries exclusives au sein du catalogue de Netflix et de comprendre si la plateforme met en avant ses propres œuvres au détriment des autres.
Ce rapport vise à détailler les différentes étapes de notre travail de récupération et analyse des données de Netflix afin de comprendre l'importance de l'exclusivité des œuvres dans l'algorithme de recommandation de la plateforme.

Outils utilisés pour notre étude

Python:

La totalité de notre programme de récupération de données est écrit en langage Python.

Selenium:

Nous avons choisi d'utiliser la librairie Selenium de Python pour automatiser les tâches répétitives de récupération de données utiles à notre étude. Ce Framework nous a été très utile car il a permis de récupérer en quelques heures près d'un milliers d'œuvres du catalogue Netflix ainsi que leurs caractéristiques telles que le genre, le lien vers la miniature… De la même manière, c'est avec Selenium que nous avons entraîné nos profils en likant et dislikant des films et séries automatiquement afin de récupérer une seconde base de données permettant de tirer des conclusions utiles après comparaison avec la première.

Gephi:

Nous avons décidé d'utiliser le logiciel Gephi pour la visualisation de nos données. Cet outil est très utile car il permet assez facilement de représenter un nombre important de données et les liens existant entre elles, très utile dans notre cas pour repérer d'éventuels clusters d'oeuvres « Netflix originals » à partir des recommandations faîtes par la plateforme après le visionnage de chaque film ou série.

Protocole

Profils

Notre étude consistera en une récupération des données de trois profils nouvellement créés et leur analyse, ce qui devrait nous permettre de tirer des conclusions intéressantes quant à la volonté qu'a Netflix de mettre en avant ses contenus exclusifs par le biais de son algorithme de recommandation.

Les trois profils sont définis selon le comportement arbitraire que nous avons choisi et qui dictera les actions qu'ils effectueront sur la plateforme :

- Profil témoin : ce premier profil n'effectuera pas d'actions sur le site, nous le créons simplement pour connaître la proportion d'œuvres exclusives dans le catalogue d'un nouvel utilisateur. Ce profil nous permettra notamment de comparer cette proportion avec celle que nous serions censés retrouver, à savoir la proportion d'œuvres exclusives dans le catalogue entier.

- Profil "Fan" : Ce premier profil de test est un fan absolu des films et séries produits par la plateforme. Sa démarche consiste à liker toutes les œuvres exclusives et à disliker celles qui ne le sont pas.

- Profil "Hostile" : Ce deuxième profil de test est l'antipode du "Fan". Ici, nous faisons liker au profil toutes les œuvres qui ne sont pas exclusives et disliker toutes celles qui le sont.

Nous pouvons noter que les actions réalisées par les profils consistent simplement à liker et disliker des œuvres et non à les visionner. Nous pensons que le visionnage des films et séries est inutile pour tenter de montrer à l'algorithme de recommandation de la plateforme qu'une œuvre nous a plu ou déplu. Par contre, il nous semble logique que les réactions à un film ou une série comme un like ou un dislike soient assez rares et démontrent à Netflix que l'œuvre nous a particulièrement touchée de manière positive ou négative. Il nous semble donc évident que liker ou disliker des œuvres comme nous le faisons agit sur les recommandations que Netflix va nous faire.

Récupération de données

Pour chaque profil, nous récupérons les informations de toutes les œuvres présentes sur la page principale ainsi que de celles recommandées par la plateforme pour chaque film ou série. Ce nombre peut varier suivant les profils mais il est d'environ XXX œuvres ce qui nous paraît juste pour considérer que notre échantillon est représentatif et que nos résultats seront justes. Cette récupération fera office de premier passage.

Ensuite, nous effectuerons les actions précédemment détaillées correspondant au bon profil de manière automatisée sur toutes les œuvres. Après cela, nous pourrons procéder à une deuxième récupération des données ce qui constituera le deuxième passage.

Résultats et analyse

Après avoir récupéré les données des trois profils pour le premier et deuxième passage, nous procéderons à la création de graphes Gephi pour tenter de comprendre les liens qu'ont les œuvres recommandées par Netflix entre elles. C'est ici que nous exploiterons les œuvres recommandées pour chaque œuvre que nous aurons préalablement récupérées dans notre base de données.

Résultats attendus

Les échantillons étant représentatifs car assez important en données, une certaine différence dans la proportion d'oeuvres exclusives présentes sur les catalogues de nos profils par rapport à celle dans le catalogue complet suffirait à nous faire penser que Netflix adopte à travers son algorithme de recommandation un comportement de mise en avant des films et séries exclusifs.

Création des profils

La première étape consiste à créer nos profils sur lesquels nous récupérerons les films et séries proposés lors de la création et que nous entraînerons ensuite suivant un comportement donné et détaillé dans notre protocole.

Lors de la création des profils, nous choisissons la langue « français » et nous ne sélectionnons pas de films en particulier, pour que la plateforme ne fasse aucune supposition et que les données soient les plus neutres possible. Nous considérons que ces deux choix permettent d'obtenir les données les plus neutres sans influer sur les films et séries que l'algorithme propose lors de la création du profil.

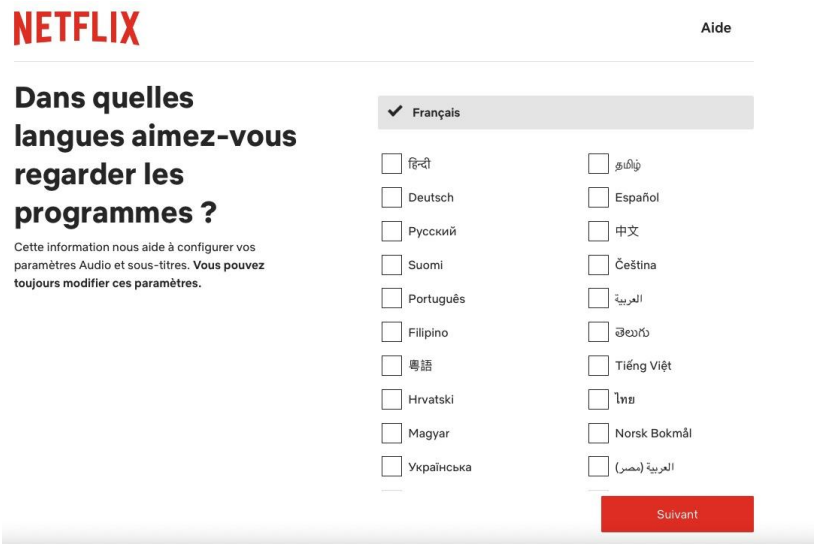

Figure 1 - Choix de la langue lors de la création des profils

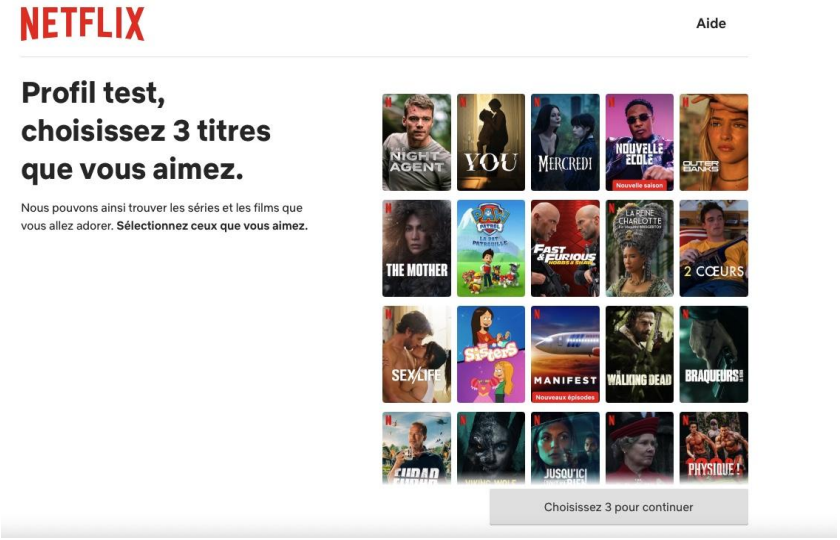

Figure 2 - Sélection des films lors de la création des profils

Récupération de données

Après la création des profils, nous devons récupérer l'intégralité des XXX œuvres présentes sur les pages d'accueil et les inscrire dans trois fichiers .csv qui constitueront nos bases de données. L'opération dure quelques heures ce qui peut paraître assez long, nous avons ici préféré la quantité de données à la rapidité de l'action, puisque nous avons recueilli des données que nous n'utiliserons finalement pas comme le réalisateur ou encore les actrices

et acteurs pour obtenir des base de données riches et exploitables pour de potentielles études futures. Ce scrapping nous permet d'avoir en notre possession des bases de données utiles pour de prochaines potentielles études sur la plateforme. Nous avons aussi décidé de noter dans la base de données le nombre d'occurrences des œuvres pour se rendre compte de l'importance et de la tendance qu'a Netflix de les mettre en avant.

Tout le travail de récupération de données est automatisé, nous demandons dans notre programme de cliquer sur une œuvre de la page d'accueil puis de récupérer toutes les informations que nous souhaitons obtenir à partir du HTML/CSS de la page, et ce pour les XXX œuvres de la page.

Vérification de l'exclusivité d'une oeuvre

Maintenant que nous avons en notre possession des bases de données intéressantes de quelques centaines d'œuvres qui correspondent au premier passage, nous pouvons comme nous l'avons souligné dans le protocole rajouter dans nos fichiers une colonne qui assigne la valeur « Vrai » lorsque l'œuvre est exclusive à la plateforme et « Faux » dans le cas contraire. Nous remarquons rapidement qu'il nous est impossible d'utiliser le HTML/CSS de la page comme nous l'avions fait jusqu'à maintenant car aucune information concernant l'exclusivité du contenu n'y est inscrite. Nous remarquons alors que l'unique différence entre les œuvres exclusives et celles qui ne le sont pas réside dans leur miniature. En effet, un logo est présent sur celles qui le sont et il n'est pas présent sur les autres. Nous adaptons ainsi notre programme pour qu'il récupère le lien de la miniature d'une œuvre lors de la récupération de ses données et compare certains pixels en haut à gauche avec un code RGB qui nous sert de témoin et que nous avons préalablement récupéré à partir du logo Netflix lui-même. Cette solution algorithmique nous permet de gagner beaucoup de temps puisqu'elle se déploie en même temps que la collecte des données et ne nécessite ainsi pas de temps supplémentaire.

Après plusieurs dizaines de tests faits « à la main », nous lançons cette partie du programme et remarquons que la solution algorithmique nous donne le résultat juste à chaque fois. Au vu de son exactitude sur un nombre déjà important d'œuvres, nous décidons de considérer que nous pouvons faire confiance à cette solution et nous l'utilisons pour toutes nos œuvres.

Entraînement du profil

Avec en notre possession des bases de données intéressantes de plusieurs centaines d'œuvres, nous pouvons procéder à l'entraînement des profils avant de faire une deuxième récupération de données qui correspondra au deuxième passage.

Pour ce faire, nous prenons la liste de toutes les œuvres présentes dans la base de données du premier passage du profil correspondant et nous likons ou dislikons l'œuvre suivant le profil choisi et son comportement associé.D'un point de vue algorithmique, cette étape consiste à rentrer dans le navigateur l'url de la fiche d'informations des oeuvres

préalablement récupéré lors de la collecte de données avant de cliquer sur le bouton like ou dislike. Selenium nous permet d'automatiser cette étape et de cliquer sur la réaction choisie en fonction de la valeur du booléen de l'attribut "exclusif à Netflix " de l'œuvre dans la base de données. Notons que nous ne réalisons aucune action pour le profil témoin, nous ne faison que cliquer sur la fiche d'informations des oeuvres sans exprimer de réactions à leur égard.